EVALUACIÓN DE MOTORES DE BÚSQUEDA DE IMÁGENES EN LA WORLD WIDE WEB

Carlos Leal Zubiete
Universidad Carlos III de Madrid

-

INTRODUCCIÓN

Desde mediados de los años 70 en el ámbito de la documentación se viene discutiendo el problema de la explosión informativa, un fenómeno que ha tomado dimensiones dramáticas con el desarrollo a partir de los años 90 de la *World Wide Web.* Así, cuando en 1995 surgió en Dublín (Ohio) el estándar de metadatos Dublin Core, la web contaba aproximadamente con medio millón de documentos; hoy se cuentan por miles de millones.

Este problema se ha trasladado en buena medida al mundo de la imagen como consecuencia del desarrollo de las tecnologías digitales, que permiten producir tanto imágenes como documentos de audio y vídeo de una forma barata y sencilla, los cuales posteriormente pueden distribuirse a través de Internet y hacerse así accesibles a una audiencia global. Sin embargo, este crecimiento exponencial en la cantidad de imágenes (fijas y en movimiento) disponibles para el usuario carece de valor si no existen procedimientos para que su contenido sea descubierto y utilizado.
En el mundo textual, existen numerosos mecanismos de indización automatizada y algoritmos de recuperación que producen resultados aceptables, y que son el fundamento de los motores de búsqueda populares como Yahoo o Google. Por el contrario, por su propia naturaleza las imágenes son menos permeables a la descripción y la indización automatizada, lo que dificulta en buena medida su recuperación. Además, resulta evidente que los esfuerzos que los grandes motores de búsqueda están haciendo en este ámbito no son tan intensos como los que se dedican a la recuperación textual.

En este artículo evaluaremos los tres principales servicios de recuperación de imágenes disponibles en la *World Wide Web:* los pertenecientes a los dos grandes motores de búsqueda Google y Yahoo y el buscador específico Picsearch. Para ello, en primer lugar estudiaremos cuáles son las principales tendencias de investigación en recuperación de imágenes y qué retos plantea Internet. Posteriormente haremos un análisis comparativo de las características fundamentales proporcionadas por cada uno de los servicios, y finalmente tomaremos medidas experimentales de precisión en la recuperación.

LÍNEAS DE INVESTIGACIÓN EN LA RECUPERACIÓN DE IMÁGENES

B.J. Wielinga *et al.* (2001) analizan diversos paradigmas para la recuperación de imágenes fijas, agrupándolos en torno a dos categorías fundamentales:

• **Recuperación de imágenes basada en contenido** *{Contení Based Image Retrievat):*

Desarrollada fundamentalmente por ingenieros de *software,* esta línea de investigación se basa en el uso de algoritmos de indización que extraen directamente de las imágenes información relacionada con colores, texturas, formas... Dicha información se almacena posteriormente en metadatos y se utiliza para la recuperación. Generalmente el interfaz de búsqueda suele permitir recuperar imágenes similares a una proporcionada o seleccionada por el usuario (p. ej. el sistema Blobworld[1], desarrollado por la Universidad de Berkeley), o bien a un boceto creado por éste sobre la marcha (p. ej. la tecnología de IBM QBIC, utilizada por el buscador de imágenes del museo L'Hermitage[2]). Remco Veltkamp y Mírela Tañase (2002) proporcionan una panorámica muy completa de los sistemas CBIR.

En la actualidad, muchos de los esfuerzos desarrollados en esta línea se concentran en torno al estándar MPEG-7 *{Multimedia Contení Description Interface),* aprobado como norma ISO/IEC 15938/2002.

• **Recuperación de imágenes basada en texto** *(Text Based Image Retrieval):* En este caso, la recuperación se basa en metadatos textuales que describen el contenido de la imagen, ya sea a través de lenguaje natural, palabras clave o descriptores extraídos de un vocabulario controlado. Estándares genéricos como *Dublin Core* pueden asociarse a las imágenes de un modo efectivo para facilitar la recuperación, utilizando para ello un software como RDFPIC desarrollado por el *World Wide Web Consortium.* En el lado más elevado del espectro quedan las técnicas de representación del conocimiento asociadas a la Web Semántica, como las ontologías o los *Topic Maps,* que permiten una descripción compleja de la imagen descomponiéndola en estructuras.

No obstante, es importante reseñar que las técnicas referidas en este capítulo son esencialmente experimentales, y su aplicación extensiva en la World Wide Web resulta compleja. En el próximo apartado analizaremos cuáles son las técnicas efectivas que emplean los motores de búsqueda en Internet para recuperar imágenes.

RECUPERACIÓN DE IMÁGENES EN LA *WORLD WIDE WEB*

Aunque las técnicas apuntadas en el apartado anterior resultan a primera vista muy prometedoras, cuando se desciende a la realidad de la *World Wide Web* el punto de vista cambia. La gran mayoría de los usuarios de Internet está acostumbrada a trabajar con interfaces textuales de interrogación, con lo que los entornos de recuperación estrictamente basados en contenido resultan en apariencia demasiado engorrosos. El usuario medio de un motor de búsqueda espera encontrar una casilla de formulario en la que describir el contenido que se desea recuperar, sin tener que aportar una imagen previa ni mucho menos realizar un boceto. Por su parte, los sistemas de identificación automática de formas (que se aplican por ejemplo al reconocimiento facial en fotografías) están aún en fase experimental y su efectividad es limitada.

En cuanto a la posibilidad de usar para la recuperación los metadatos asociados por sus creadores a las imágenes, en la actualidad es una utopía. Sólo una mínima parte de los elementos gráficos disponibles en la web tienen asociados metadatos, con calidad diversa, y asignarlos manualmente desde la base de datos central no es factible dado el enorme volumen de información que sería necesario procesar. Así pues, se impone la necesidad de encontrar otras soluciones más prácticas.

Desgraciadamente, los motores de búsqueda son muy celosos con la información sobre el funcionamiento de sus sistemas, por lo que sólo se puede conjeturar sobre las técnicas específicas que utilizan para la recuperación de imágenes[3]. En cualquier caso, resulta evidente que emplean un modelo mixto que combina elementos de CBIR y TBIR.

Por medio del análisis automatizado de imágenes, los motores de búsqueda almacenan metadatos técnicos que les permiten distinguir entre fotografías y gráficos, color y blanco y negro... También recogen datos básicos de la imagen como el formato en el que está almacenada (JPG, GIF o PNG), sus dimensiones en pixels...

Un uso específico de los algoritmos de indización automatizada es el que permite identificar las imágenes analizadas que son iguales (i.e. un mismo elemento gráfico presente en varios sitios web) o muy similares (por ejemplo, varias versiones de *La Gioconda,* de Leonardo Da Vinci). Esto resulta especialmente útil a la hora de recuperar imágenes, ya que limita el número de resultados redundantes

que se ofrecen al usuario; además, la aparición en diversos sitios web puede ser un criterio relevante para jerarquizar los resultados de una búsqueda.

En todo caso, el mayor esfuerzo que deben hacer los motores de búsqueda al indizar imágenes es almacenar automáticamente los elementos textuales que describan el contenido. Para ello hay diversas zonas de extracción preferentes, cada una con sus ventajas y sus inconvenientes:

• **El nombre del archivo:** Suele proporcionar información certera sobre el contenido de la imagen; así, un archivo llamado "gioconda.jpg" probablemente será una reproducción del cuadro. El principal problema con esta técnica es que la mayor parte de las imágenes publicadas en la web, particularmente en los sitios comerciales importantes, suelen tener nombres tan poco expresivos como "l.jpg".

• **El atributo "ALT":** Las *Web Accesibility Guidelines** prescriben que todos los elementos gráficos de una página web deben tener un texto alternativo, que oriente a los invidentes que usan lectores de pantalla y a los usuarios que emplean navegadores sólo textuales. No obstante, las reglas destacan que en el atributo "ALT" no debe figurar una simple descripción de la imagen, sino un equivalente textual de su función en la página. Así, por ejemplo, en el caso de un logotipo de empresa que se use como vínculo a la página principal del sitio web, el texto alternativo debería ser "ir a la página principal", y no "logotipo de la empresa".

• **Los atributos "TITLE" y "LONGDESC":** Los atributos "TITLE" y "LONGDESC" de **la** etiqueta "IMG" sí son más indicados para describir el contenido de los elementos gráficos. Sin embargo, su uso es escaso dentro de la *World Wide Web.*

• **El título de la página o el texto que rodea la imagen:** Cuando no hay otra forma de obtener información sobre el contenido de la imagen, es necesario recurrir o bien al título de la página (asumiendo que éste recoge el tema principal, al que también hacen referencia las imágenes), o bien al texto que rodea a la etiqueta "IMG", que puede ser muy relevante en las páginas que ponen a continuación un pie de foto (por ejemplo, algunos medios de comunicación).

En todo caso, resulta evidente que todas estas técnicas son de por sí imprecisas, por lo que sólo combinando los algoritmos de indización automatizada de imágenes con una extracción ponderada de información textual se puede lograr que un motor de búsqueda obtenga resultados pertinentes a partir de un repositorio masivo.

PRESENTACIÓN DE LOS MOTORES DE BÚSQUEDA OBJETO DE ESTUDIO

El objetivo de este estudio es evaluar el funcionamiento de los principales servicios de recuperación de imágenes disponibles en la *World Wide Web.* A falta de datos específicos sobre este ámbito, partiremos de las cifras generales sobre audiencia de los motores de búsqueda. Según datos de Nielsen NetRatings (actualizados en marzo de 2005), Google es el gran dominador de las búsquedas en Estados Unidos con un 47.3% de consultas, seguido por Yahoo, MSN, AOL y Ask Jeeves[5]. Un estudio publicado en enero de 2005 por Keynote Systems[6] ofrece datos similares a escala global; el buscador más usado por los usuarios sigue siendo Google, si bien el motor de búsqueda de Yahoo, MSN, Ask Jeeves y Lycos están reduciendo la distancia.

Sin embargo, a pesar de la aparente variedad del panorama presentado, en la práctica existen únicamente tres grandes bases de datos de imágenes: las de Google y Yahoo/Overture y la del buscador específico Picsearch. El resto de motores de búsqueda o bien no ofrecen la posibilidad de recupe-

rar imágenes o dependen de bases de datos externas, a las que en ocasiones aplican sus propios algoritmos de ordenación de resultados. La Tabla 1 resume la situación actual de los principales servicios de búsqueda de imágenes en la *World Wide Web.*

Así pues, para la evaluación tendremos en cuenta únicamente las tres bases de datos de imágenes citadas, consultándolas a través de sus interfaces principales. En primer lugar presentaremos cada uno de los motores de búsqueda que se van a analizar. Posteriormente recopilaremos información sobre los factores físicos que inciden en la recuperación, partiendo de las categorías establecidas por Oppenheim et al. (2000). Por último, tomaremos medidas de relevancia formulando una serie de consultas a cada motor de búsqueda.

BB.DD. de imágenes	**Sitio**	**Algoritmo de *ranking***
Google	Google	Propio
	AOL	Igual que Google
Yahoo/Overture	Yahoo Search	Propio
	Alta vista	Igual que Yahoo Search
	Alltheweb	Igual que Yahoo Search
Picsearch	Picsearch	Propio
	Ask Jeeves	Propio
	MSN	**Igual que Picsearch**
	Lycos	Igual que Picsearch

Tabla 1. Servicios de recuperación de imágenes en la WWW

1. Google

Creado por Larry Page y Sergey Brin mientras estudiaban en la Universidad de Stanford a finales de los años 90, con el paso del tiempo el motor de búsqueda Google se ha convertido en la opción preferida de una gran mayoría de los internautas. En la actualidad, dice contener en su base de datos más de ocho mil millones de páginas web, si bien algunos estudios cuantitativos ponen en duda esta cifra[7]. Su servicio de recuperación de imágenes es igualmente pionero en la *World Wide Web;* la versión principal está disponible desde junio de 2001.

2. Yahoo

Fundado en 1995, cuando la *World Wide Web* era aún un fenómeno incipiente, Yahoo ganó fama mundial rápidamente por su directorio, una colección jerarquizada de vínculos a recursos disponibles en Internet alimentada manualmente. Para completar los resultados de las búsquedas que no es capaz de resolver su propio directorio, Yahoo firmó sucesivos acuerdos con los motores de búsqueda Open Text (1995-1996), Altavista (1996-1998), Inktomi (1998-2000) y Google (2000-2004). Sin embargo, tras comprar en el año 2003 tanto Inktomi como Overture, propietaria a su vez de Altavista y Alltheweb, la compañía decidió desarrollar su propio motor de búsqueda para hacer competencia directa a Google. Así pues, en la actualidad Yahoo cuenta con tres modelos de negocio. Su página principal (http://www.yahoo.com/) es un portal horizontal que aglutina servicios como noticias, correo electrónico, alojamiento de páginas web, etc.; por su parte, el directorio de recursos está alojado en http://dir.yahoo.com/; por último, al motor de búsqueda se puede acceder directamente a través de la URL http://search.yahoo.com/. Yahoo Search nació en febrero de 2004, con una base de datos cuyo tamaño se estima similar a Google. Apenas un mes después lanzó su propio servicio de búsqueda de imágenes, y en la actualidad también permite recuperar archivos de vídeo.

3. **Picsearch**

Fundado en Estocolmo en el año 2000, Picsearch es el principal motor de búsqueda específico de imágenes disponible en la *World Wide Web*. Su importancia en este campo se ve realzada por el amplio número de buscadores que recurren a su base de datos para nutrir sus servicios de búsqueda de imágenes, entre los que destacan MSN, Ask Jeeves y Lycos, los tres principales rivales de Google y Yahoo Search.

EVALUACIÓN DE LOS MOTORES DE BÚSQUEDA

1. Factores explícitos

El primer paso dentro de la evaluación de los servicios de recuperación de imágenes de Google, Yahoo y Picsearch pasa por atender a lo que Martínez Méndez y Rodríguez Muñoz (2002) definen como aspectos explícitos del motor de búsqueda; es decir, aquellos factores externos y evidentes que afectan a la utilidad del sistema. Como guía en el análisis emplearemos los once criterios señalados por Oppenheim, Morris y McfCnight (2000) para la evaluación de motores de búsqueda, adaptándolos a las especificidades de la recuperación de imágenes.

a) Tamaño de la base de datos

Al contrario de lo que sucede con los motores de búsqueda textuales, en el ámbito de la recuperación de imágenes no hay estudios cuantitativos fiables que den cuenta del tamaño real de las bases de datos para la recuperación de imágenes. Ateniéndonos a sus propios datos, Google recoge más de 1.150 millones de elementos gráficos. Por su parte, Yahoo Search afirma realizar sus búsquedas sobre más de 1.500 millones de imágenes. Actualmente Picsearch no ofrece datos sobre su cobertura, si bien resulta evidente que es muy inferior a la manejada por Google o Yahoo[8]. Según Ask Jeeves, que emplea la base de datos de Picsearch para sus búsquedas, la cifra ronda los 500 millones de imágenes.

b) Actualización del índice

En la documentación de ayuda que proporcionan a los usuarios, los tres motores de búsqueda objeto de análisis señalan que la actualización de la base de datos se produce diariamente. No obstante, haciendo búsquedas sobre imágenes de la actualidad informativa se comprueba que Yahoo Search y Picsearch ofrecen los mejores resultados, mientras que Google tarda notablemente más en incorporar nuevas fotografías. De hecho, a lo largo del año 2004 Google Images sufrió serios problemas de actualización que se prolongaron varios meses, pero que ya están solucionados[9].

c) Relevancia

Para muchos estudios, la relevancia de los resultados recuperados es el principal factor, cuando no el único, que debe tenerse en cuenta al evaluar un motor de búsqueda. El principal problema con la relevancia es que con frecuencia depende de la percepción del usuario final; así, con frecuencia ante una misma consulta un usuario especializado en el tema sobre el que se interroga al sistema no considerará relevantes los mismos resultados que alguien cuyo conocimiento previo de la materia sea menor. En todo caso, la búsqueda de imágenes facilita el establecimiento de criterios que permitan discernir claramente entre resultados relevantes y no relevantes. Con el fin de completar este estudio, se han tomado medidas de relevancia cuyos resultados están disponibles en el apartado 2.

La posibilidad de emplear operadores booleanos y de proximidad en las consultas facilita la formulación de estrategias de búsqueda complejas, que proporcionan resultados más precisos. Dentro de los motores de búsqueda de imágenes, tanto Yahoo Search como Google permiten el uso de los tres operadores booleanos (AND, OR y NOT), mientras que Picsearch únicamente soporta AND (por defecto) y NOT. Además, los tres permiten el uso de operadores de adyacencia (comillas), pero no de proximidad.

e) Materias

Oppenheim, Morris y McKnight (2000) señalan que no todos los motores de búsqueda tienen por qué tener el mismo rendimiento en distintas áreas temáticas. En ese sentido, a la hora de evaluar globalmente un motor de búsqueda es recomendable tener en cuenta que si se trabaja con una sola área temática los resultados extraídos pueden no ser extrapolables al conjunto.

f) La dinámica naturaleza de la web

Al contrario de lo que sucede con las bases de datos tradicionales, la *World Wide Web* es una I realidad cambiante, en la que los contenidos aparecen y desaparecen a gran velocidad. En ese sentido, los motores de búsqueda de imágenes deben establecer mecanismos de garantía que comprueben periódicamente el estado de los recursos almacenados para minimizar la aparición de "errores 400". En el apartado 2 expondremos los resultados de las medidas porcentuales tomadas en torno a la cantidad de vínculos desactualizados en cada motor de búsqueda.

g) **Tiempo de respuesta**

Dimensión clásica en los estudios experimentales de sistemas de recuperación de información, en la web resulta difícil de controlar y poco significativa, por cuanto que en la mayoría de los casos el tiempo de respuesta del sistema se mide en décimas o centésimas de segundo.

h) Características diferentes del sistema

En los primeros tiempos de la *World Wide Web* los distintos sistemas de recuperación de información ofrecían características muy diversas, que podían ayudar a los usuarios a decantarse por uno u otro. Sin embargo, tras el éxito de Google los servicios ofrecidos por los motores de búsqueda se han estandararizado en buena medida, y sólo pequeños detalles diferencian a unos sistemas de otros. Así, por ejemplo, Google y Yahoo ofrecen sugerencias para la corrección de erratas en las búsquedas, una característica no disponible en Picsearch. Probablemente la característica individual más destacada sea la que ofrece la base de datos de Yahoo a través del interfaz de Altavista: a través del vínculo "más información" en la página de resultados permite acceder a los metadatos almacenados sobre la imagen, incluyendo los términos seleccionados en la página de origen y otras direcciones que también contienen el elemento gráfico.

i) Opciones de búsqueda

Contar con filtros avanzados para limitar los resultados de la búsqueda puede ser una ayuda fundamental para realizar tareas de recuperación complejas. Así, casi todos los motores de búsqueda permiten limitar las búsquedas por tamaño o por si la imagen es en color o en blanco y negro. Un uso avanzado de técnicas CBIR permite distinguir entre fotografías y gráficos, o entre imágenes estáticas y animaciones. La Tabla 2 recoge las opciones de búsqueda avanzadas que soporta cada uno de los motores objeto de estudio.

	Google	Yahoo	Picsearch
Tamaño	•	«•	<•
Color / ByN	>•	<y	•
Animación/imagen fija	X	X	•
Formato de imagen	•	**X**	X
Fotografía/elemento gráfico	X	•[1]	X
Sitio/dominio de origen	V*	«•	X

[1] Sólo disponible a través del interfaz de Altavista

Tabla 2. Filtros aplicables en motores de búsqueda de imágenes

j) Factores humanos y cuestiones de la interface

Los estudios sobre usabilidad y accesibilidad de sitios web tienen un amplio desarrollo, y prácticamente no hay ningún trabajo de evaluación global de motores de búsqueda que no tenga en cuenta factores de este tipo. Sin embargo, como Oppenheim, Morris y McKnight (2000) explican, este tipo de consideraciones tienen dos problemas fundamentales: el componente subjetivo que suelen acarrear los estudios de usabilidad y la rápida evolución de los interfaces, que además tienden cada vez más a la estandarización. Para usuarios de lengua no inglesa, en todo caso, es interesante señalar que tanto Google como Yahoo ofrecen el interfaz traducido a un buen número de idiomas, incluido el español, mientras que Picsearch sólo está disponible en inglés, alemán, danés, noruego y sueco.

k) Calidad de los resúmenes

Con el objeto de ayudar a los usuarios a seleccionar rápidamente a qué imágenes les interesa acceder, los motores de búsqueda almacenan versiones reducidas de las imágenes, conocidas como thumbnails. Esta práctica no ha estado exenta de una cierta polémica, ya que para algunos propietarios de imágenes supone una violación de sus derechos de propiedad intelectual; sin embargo, por el momento la mayor parte de los pleitos se han resuelto favorablemente a los intereses de los motores de búsqueda[10]. Además de las vistas en miniatura, en los tres casos el interfaz ofrece información sobre las dimensiones de la imagen en pixels, su tamaño en bytes y el formato de codificación.

2 Medidas de relevancia

En paralelo a los factores físicos analizados en el apartado anterior, con el fin de evaluar globalmente los motores de recuperación de imágenes objeto de nuestro estudio hemos descendido al plano experimental para tomar medidas de relevancia. Para ello, hemos formulado diez consultas simples que cubren ámbitos temáticos diversos (cine, deporte, arte, política, acontecimientos...) emplean-

do para ello una sintaxis interpretable por los tres buscadores que combina operadores de intersección (AND) y adyacencia.

Una vez formulada la consulta a cada uno de los motores de búsqueda, se han analizado los primeros veinte resultados recuperados. Para ello, en primer lugar se han visitado los 20 primeros vínculos para comprobar cuántos de ellos no están ya disponibles". Posteriormente, entre los resultados válidos se han separado aquéllos que son relevantes a la consulta formulada. En el caso de resultados duplicados, se han considerado relevantes sólo la primera vez que aparecen. Además, se han tomado medidas del número de resultados duplicados y de la cantidad de primeros aciertos (registros que aparecen en primer lugar y que están disponibles y son relevantes). La Tabla 3 muestra un resumen de los resultados obtenidos con fecha 24-V-2005, que están disponibles con más detalle en el Anexo I.

Motor	Media result.	Precisión	Res. Duplicados	Res. Erróneos	Primer acierto
Google	798,5	**63%**	**5,5%**	**2%**	**90%**
Yahoo	1914,4	**67%**	**5,5%**	**6,5%**	**90%**
Picsearch	211,1	**55%**	**2%**	**18%**	**60%**

Tabla 3. Resultados de las medidas de precisión tomadas

Lo primero que llama la atención de los resultados del trabajo práctico es que, a pesar de que el tamaño alegado de sus bases de datos es similar, Yahoo Search recupera un 140% más de resultados de media que Google sobre las diez consultas formuladas. Además, la precisión de Yahoo es ligeramente I superior (67% frente a 63%), si bien es cierto que Google es más estable en la recuperación; en las cues- | tiones formuladas sólo en una ocasión baja del 50% de precisión, frente a las dos de Yahoo Search. Esto se debe sobre todo a que Google agrupa los resultados provenientes de un mismo sitio web y presenta únicamente dos de ellos; de esta forma, limita la posibilidad de que se acumulen entre los primeros resultados registros no relevantes o ya no disponibles. Por el contrario, al permitir que entre los primeros resultados coincidan varias imágenes de un mismo sitio Yahoo consigue aumentar la relevancia pero también se vuelve más irregular puntualmente. En cuanto a la estabilidad de la base de datos, Google se muestra netamente superior con apenas un 2% de resultados erróneos frente al 6,5% de Yahoo.

Por su parte, Picsearch obtiene resultados inferiores a sus competidores tanto en precisión (apenas llega al 55%) como, sobre todo, en número de registros recuperados -211 de media en las diez consultas-. De hecho, en la consulta número 5 apenas ofrece 13 resultados frente a los 271 de Google y los 280 de Yahoo, factor que ha sido necesario tener en cuenta en los cálculos de precisión. En todo caso, el dato más significativo es el 18% de resultados erróneos, que son síntoma de una base de datos poco actualizada. Al igual que Yahoo, Picsearch permite que las imágenes de un mismo sitio web se agrupen entre los primeros veinte resultados, lo que da cuenta de los éxitos puntuales en la recuperación; sin embargo, en cuatro de las diez consultas no alcanza el 50% de precisión. Además, el porcentaje de éxito en el primer resultado es del 60%, frente al 90% de Yahoo y Google.

CONCLUSIONES

El resultado de las medidas realizadas sobre los tres servicios de búsqueda de imágenes que han sido objeto de este estudio ofrece un panorama nítido que muestra dos motores de búsqueda líderes (Google y Yahoo) y un tercero más rezagado, Picsearch, que no puede competir ni en infraestructura de búsqueda ni en precisión de los resultados con los dos anteriores.

La experiencia práctica demuestra que Yahoo aventaja ligeramente a Google no sólo en la precisión sino también en la cantidad de resultados, gracias al mayor tamaño de su base de datos, así como en actualización. Sin embargo, a pesar de tener el mismo sustrato tecnológico, paradójicamente para consultar esta base de datos es preferible emplear el interfaz de Altavista, que ofrece un mejor soporte de todos los operadores booleanos (incluyendo los paréntesis, indispensables para formular una consulta compleja), permite acceder a los metadatos almacenados de la imagen, incluyendo direcciones alternativas donde puede hallarse, y proporciona opciones más avanzadas para limitar la búsqueda.

El buscador temático Picsearch comparativamente es el peor situado de los tres. En su contra tiene no sólo el escaso tamaño de su base de datos y la menor relevancia de sus resultados, sino también cuestiones físicas como el deficiente soporte que da a los operadores booleanos (no permite usar el OR) o la escasez de opciones avanzadas de búsqueda. Además, revisando las interfaces de recuperación de imágenes de MSN, Ask Jeeves y Lycos se hace evidente que estos motores de búsqueda, que en el ámbito textual pretenden hacer competencia a Google y Yahoo, no tienen tanto interés por lo gráfico; así, en la mayoría de los casos no hay ayuda contextualizada, y ninguno de ellos permite realizar una búsqueda avanzada.

BIBLIOGRAFÍA

CHISHOLM, W.; VANDERHEIDEN, G; JACOBS, I. (ed). *Web contení accessibility guidelines.* World Wide Web Consortium Web Accessibility Initiative. En línea [consultado 31-V-2005]. En: <http://www.w3.org/TR/WAl-WEBCONTENT>

KEYNOTE SYSTEMS. "Yahoo! Search and MSN Search Cióse the Gap with Google*". 13 de enero de 2005. En línea [consultado 31-V-2005]. En: <http://searchenginewatch.com/searchday/article.php/2158711>

MARTÍNEZ MÉNDEZ, F.J.; RODRÍGUEZ MUÑOZ, J.V. "Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web". Information Research, Vol. 8 No. 2, 2003.

0PPENHE1M, C ; MORRIS , A. ; MCKNIGHT , C . "The Evaluation of WWW Search Engines". *Journal of Documentation* 56, 2000.

SHERMAN, C. "Google Polishes its Image". Search Engine Watch, 26 de junio de 2001. En línea [consultado 31-V-2005]. En: <http://searchenginewatch.com/searchday/article.php/2158711 >

SULLIVAN, D. "Image & Multimedia Search Complaints". Search Engine Watch, 21 de abril de 2004. En línea [consultado 31-V-2005]. En: <http://searchenginewatch.com/resources/article.php/2156521>

SULLIVAN, D . "Nielsen NetRatings Search Engine Ratings". Search Engine Watch , 22 de abril de 2005. En línea [consultado 31-V-2005]. En: <http://searchenginewatch.com/reports/article.php/2156451>

SULLIVAN, D. "5th Annual Search Engine Watch Awards". Search Engine Watch, 21 de marzo de 2005. En línea [consultado 31-V-2005]. En: <http://searchenginewatch.com/awards/article.php/3494141>

VELTKAMP, R.C. ; TAÑASE, M. "Content-Based Image Retrieval Systems: A Survey". Technical report, University of Utrecht, Dept. of Computing Science, 2000. En línea [consultado 31-V-2005]. En: I <http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey.pdf>

WIELINGA, B.J.; SCHRE1BER, A. TH.; WIELEMAKER, J.; SANDBERG, J.A.C. "From Thesaurus to Ontology". En: *Proceedings of the International Conference on Knowledge Capture.* New York: ACM Press, 2001.

ANEXO I. Resultados de las consultas

GOOGLE

Form. Booleana	Recuperados	Relevantes	Duplicados	Fallidos	Primer acierto
"Edward Hopper" painting	**354**	**11**	**5**	**0**	V
"Harrison Ford" "Han Solo"	**159**	10	**2**	**0**	«/
"Roger Federer" Wimbledon	639	**12**	**2**	**1**	•
"Eiffel tower" night	1610	17	**1**	**1**	V
"Miguel Indurain" tour	271	**13**	**1**	**0**	0/
Lincoln Statue	1730	18	**0**	**0**	<•
"Bill Clinton" "Al Gore"	295	4	**0**	**0**	X
"September 11"penjagon	1150	**13**	**0**	**0**	•
Rodin "Le Penseur"	287	**13**	**0**	**1**	•
Germán stamp	1490	**15**	**0**	**1**	«/

PICSEARCH

Form. Booleana	Recuperados	Relevan(es	Duplicados	Fallidos	Primer acierto
"Edward Hopper" painting	**42**	8	0	**11**	•
"Harrison Ford" "Han Solo"	114	**11**	0	6	X
"Roger Federer" Wimbledon	153	**11**	1	4	«•
"Eiffel tower" night	301	12	0	6	• 0
"Miguel Indurain" tour	13	6	0	3	•
Lincoln Statue	1029	18	1	0	X
"Bill Clinton" "Al Gore"	105	4	1	1	X
"September 11" pentagon	56	7	0	0	•
Rodin "Le Penseur"	35	**12**	1	4	•
Germán stamp	**263**	17	0	1	X

YAHOO

Form. booleana	Recuperados	Relevantes	Duplicados	Fallidos	Primer acierto
"Edward Hopper" painting	466	**12**	2	**3**	•
"Harrison Ford" "Han Solo"	963	14	2	**1**	•
"Roger Federer" Wimbledon	758	15	0	**2**	V
"Eiffel tower" night	2962	18	0	**2**	<•
"Miguel Indurain" tour	280	16	1	**0**	•
Lincoln Statue	2541	**14**	2	**1**	**X**
"Bill Clinton" "Al Gore"	521	**7**	0	**2**	<y
"September 11" pentagon	6543	14	1	**1**	«/
Rodin "Le Penseur"	167	**17**	2	**0**	•
Germán stamp	**3943**	**7**	1	**1**	«/

NOTAS

[1] Disponible *online* [consultado 31-V-2005]. En: <http://elib.cs.berkeley.edu/photos/blobworld/start.html>
[2] Demostración disponible *online* [consultado 31 -V-2005]. En: <http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English>
[3] Chris Sherman hace un buen acercamiento a los fundamentos técnicos de la recuperación de imágenes en Search Engine Watch. Disponible *online* [consultado 31-V-2005]. En<htípy/searchenginewatch.con V searchday/article.php/215 8711 >
[4] Las normas de accesibilidad están disponibles disponible *on Une* [consultado 31-V-2005]. En: <http://www.w3.org/TR/WAl-WEBC0NTENT/>
[5] Parte de este informe está disponible *on Une* [consultado 31-V-2005]. En: <http://searchengine-watch.com/reports/article.php/2156451>
[6] Un extracto de este estudio está disponible *on Une* [consultado 31-V-2005]. En:<htípy/www.Jceynote.com/news_events/releases_2005/05janl3.html>.
[7] Google anunció el 10 de noviembre de 2004 que su índice se había doblado, pasando de unos 4.000 millones de páginas indizadas hasta prácticamente el doble. No obstante, dado que esta información coincide en el tiempo con la salida a bolsa de la empresa y el incremento de la competencia por el lanzamiento de Yahoo Search, el dato debe ser manejado con cautela.
[8] En el apartado 4.2 se exponen los datos experimentales obtenidos, que muestran que Yahoo Search obtiene un 800% más de resultados que Picsearch.
[9] Danny Sullivan da cuenta del problema en Search Engine Watch. Disponible *online* [consultado 31-V-2005]. En: <http://searchenginewatch.com/awards/article.php/3494141 >
[10] Danny Sullivan resume la situación legal en Search Engine Watch. Disponible *online* [consultado 31 -V-2005]. En: <http://searchenginewatch.eom/resources/article.php/2156521>
[11] Se han considerado disponibles todas las imágenes que son accesibles en la dirección almacenada en el motor de búsqueda, al margen de que la página web en la que estuvieran integradas ya no exista.